美团云Docker平台

郑坤@美团云计算部 201610

纲要

- 美团云Docker平台介绍
- 平台设计
- 关键技术介绍
- 中长期技术规划

什么是美团云Docker平台?

- 基于美团云自研的Docker容器集群管理系统
- 提供Docker镜像托管、容器实例化、调度、运行、监控等功能
- 为新美大公有云与私有云提供Docker容器基础服务
- 2015年7月筹划，11月上线，稳定运行至今

业界Docker容器集群管理系统介绍

Google Kubernetes(k8s)
Google开源项目

- 容器集群管理，不限于Docker
- Feature丰富
- 生产环境使用较多

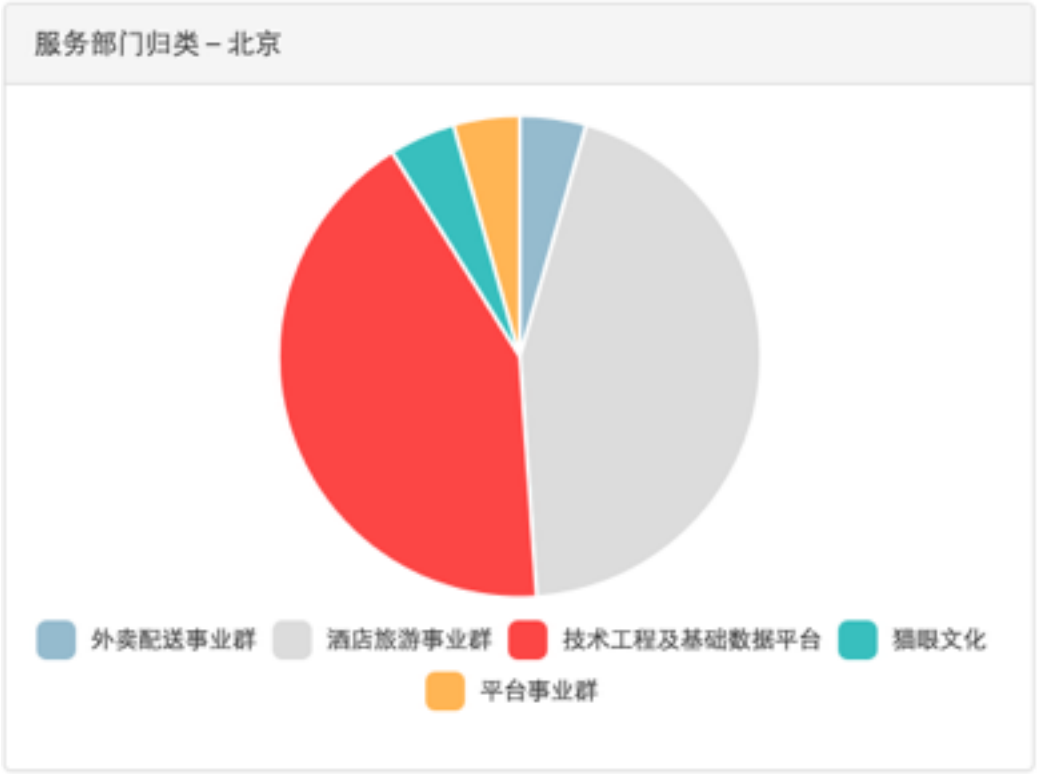
Docker Swarm
Docker开源项目

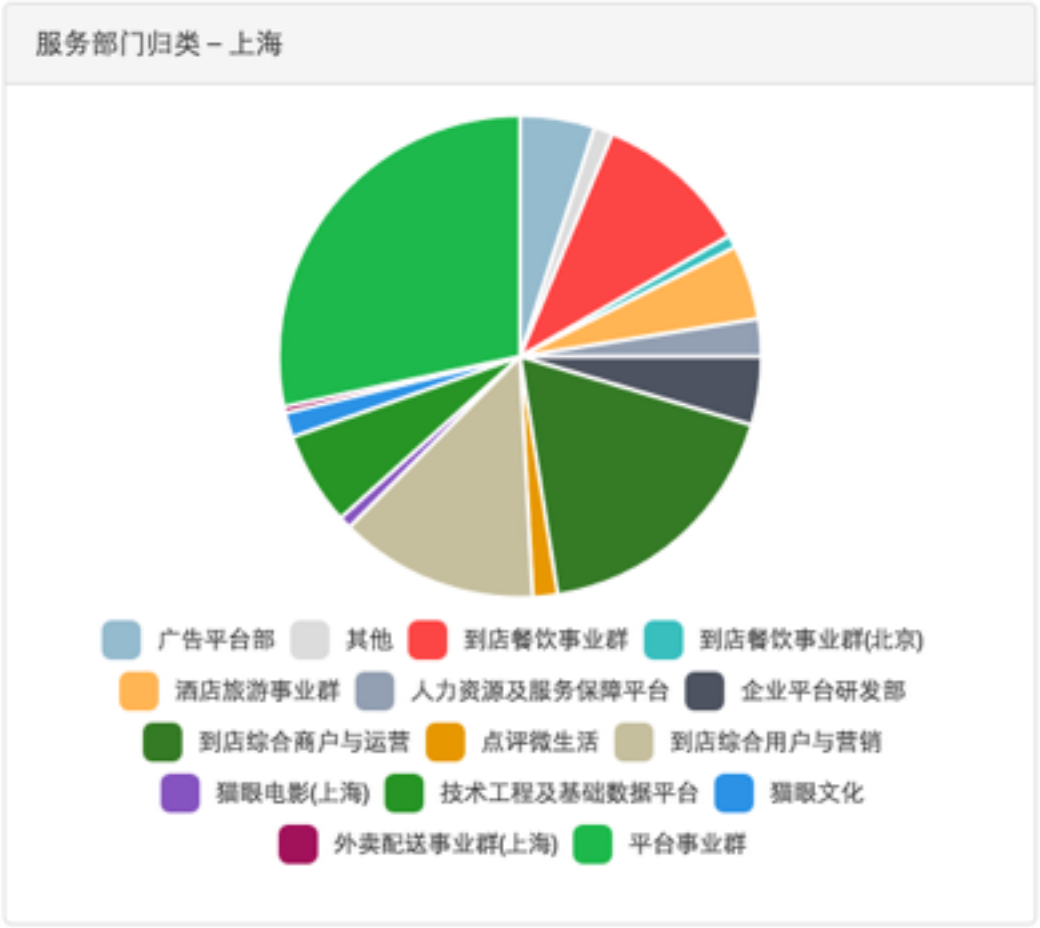
- Docker集群管理
- Feature较少
- 生产环境使用较少

为什么选择自研?

业务需求角度	为新美大业务定制系统，更接地气，能够快速满足业务需求，技术方案自主控制
工程方法论角度	复用美团云现有的轮子，兼容现有基础架构，节省运维成本，开发快

在新美大多个业务线应用





Docker平台价值

1. 提供基于容器的Devops开发模式：开发、打包、测试、部署、运维	开发高敏捷性
2. 快速故障恢复、完善的监控能力，简单高效的问题定位和诊断能力	业务高可用性
3. 支持业务达到秒级的水平和垂直伸缩	业务高弹性
4. 业务实例轻量化，免除操作系统虚拟化开销	高资源利用率

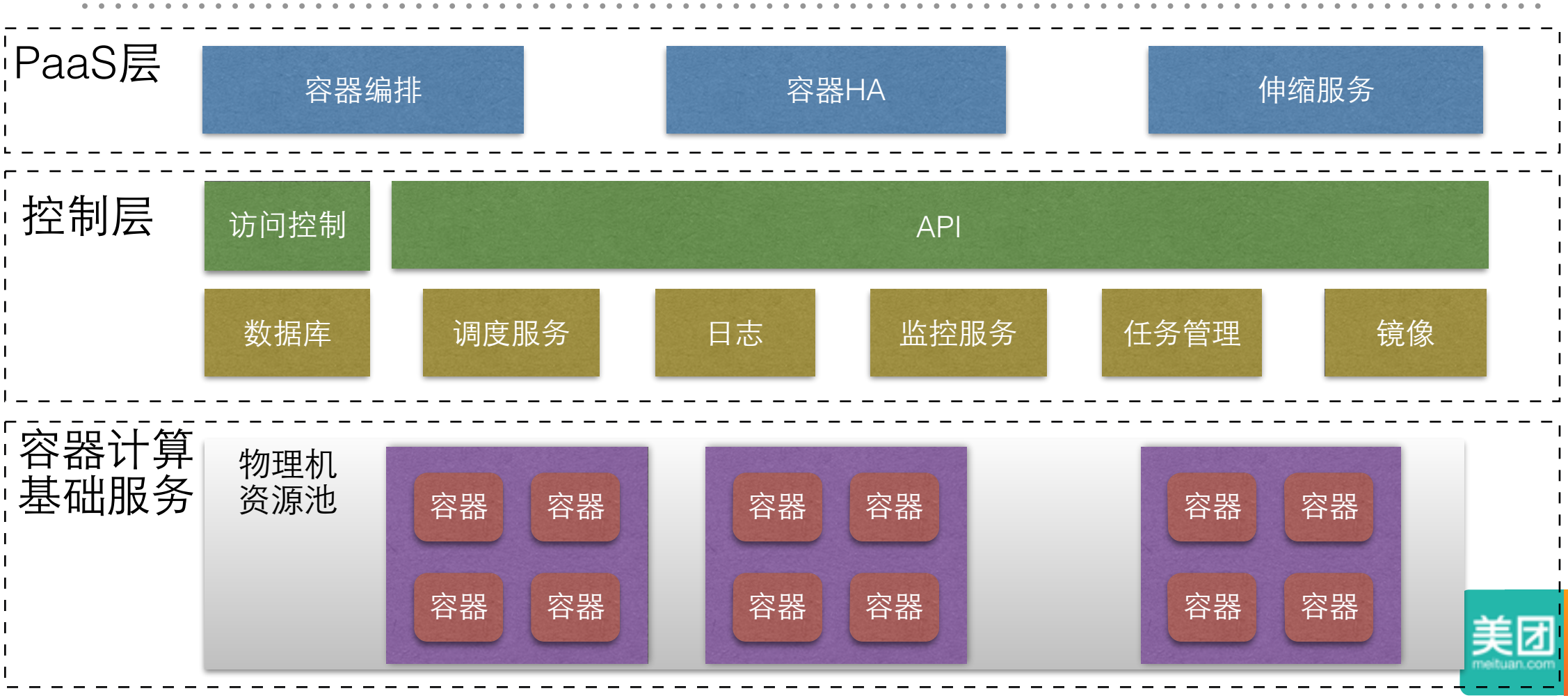
技术挑战

- 高SLA服务
  - 云计算平台级SLA服务
- 复杂的调度服务
  - 支持业务的各种调度策略，快速高效调度
- 数据高可靠
  - 数据高可靠性，具备数据备份、恢复、迁移能力
- 资源隔离
  - 多层次隔离、资源硬限制

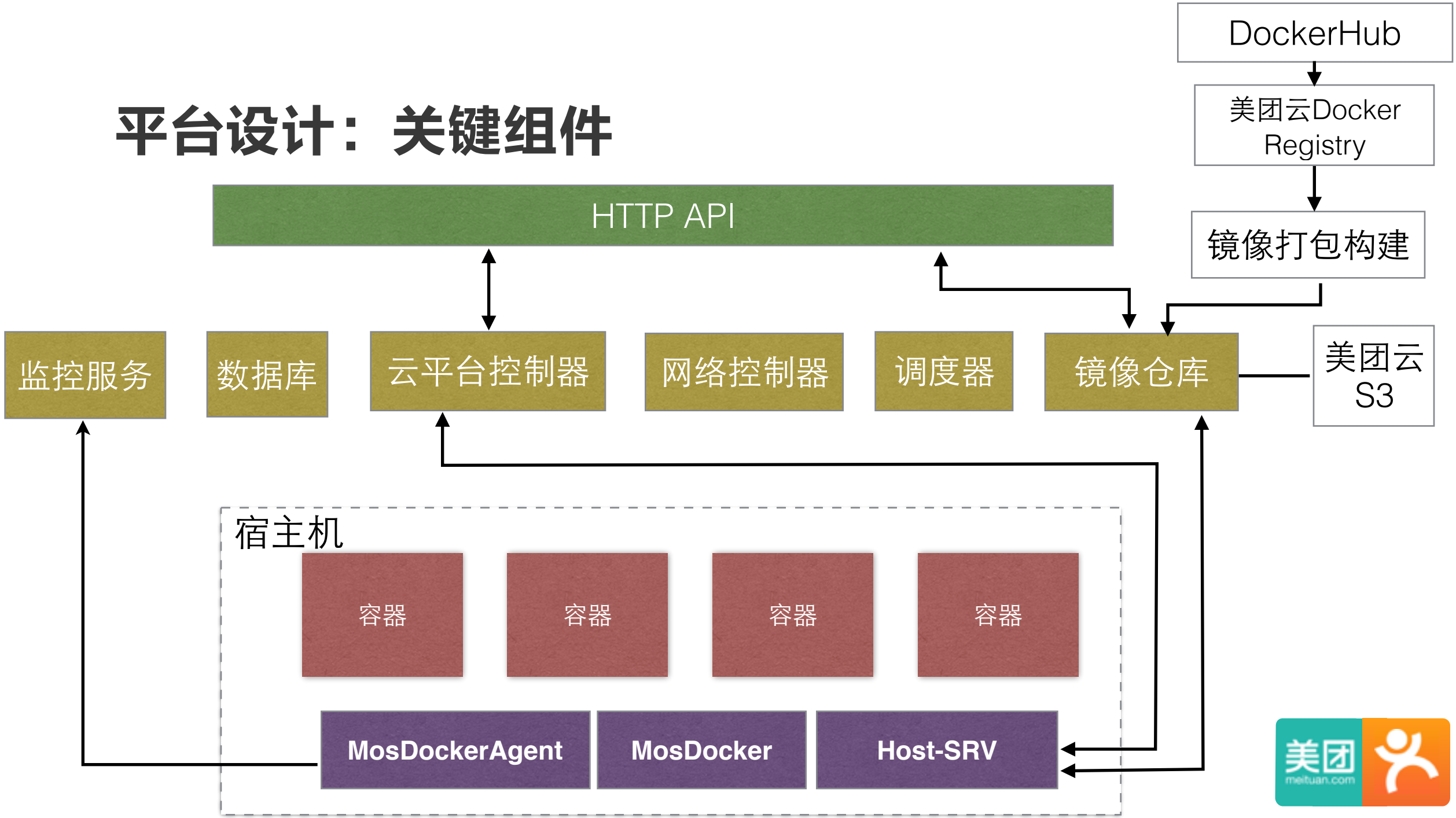
美团云Docker平台设计目标

- 可管理宿主机规模5千台，容器实例数量10万个
- SLA目标
  - 容器服务可用性99.95%
  - Volume数据可靠性99.999%
  - 秒级的容器实例伸缩能力
- 支持微服务类型业务
- 分布式网络架构、具备网络隔离能力

平台设计：逻辑架构


PaaS层
容器编排
容器HA
伸缩服务
控制层
访问控制
API
数据库
调度服务
日志
监控服务
任务管理
镜像
容器计算
基础服务
物理机
资源池
容器
容器
容器
容器
容器
容器
容器
容器
容器
容器
容器
容器

平台设计：关键组件


DockerHub
美团云Docker Registry
镜像打包构建
HTTP API
监控服务
数据库
云平台控制器
网络控制器
调度器
镜像仓库
美团云 S3
宿主机
容器
容器
容器
容器
MosDockerAgent
MosDocker
Host-SRV

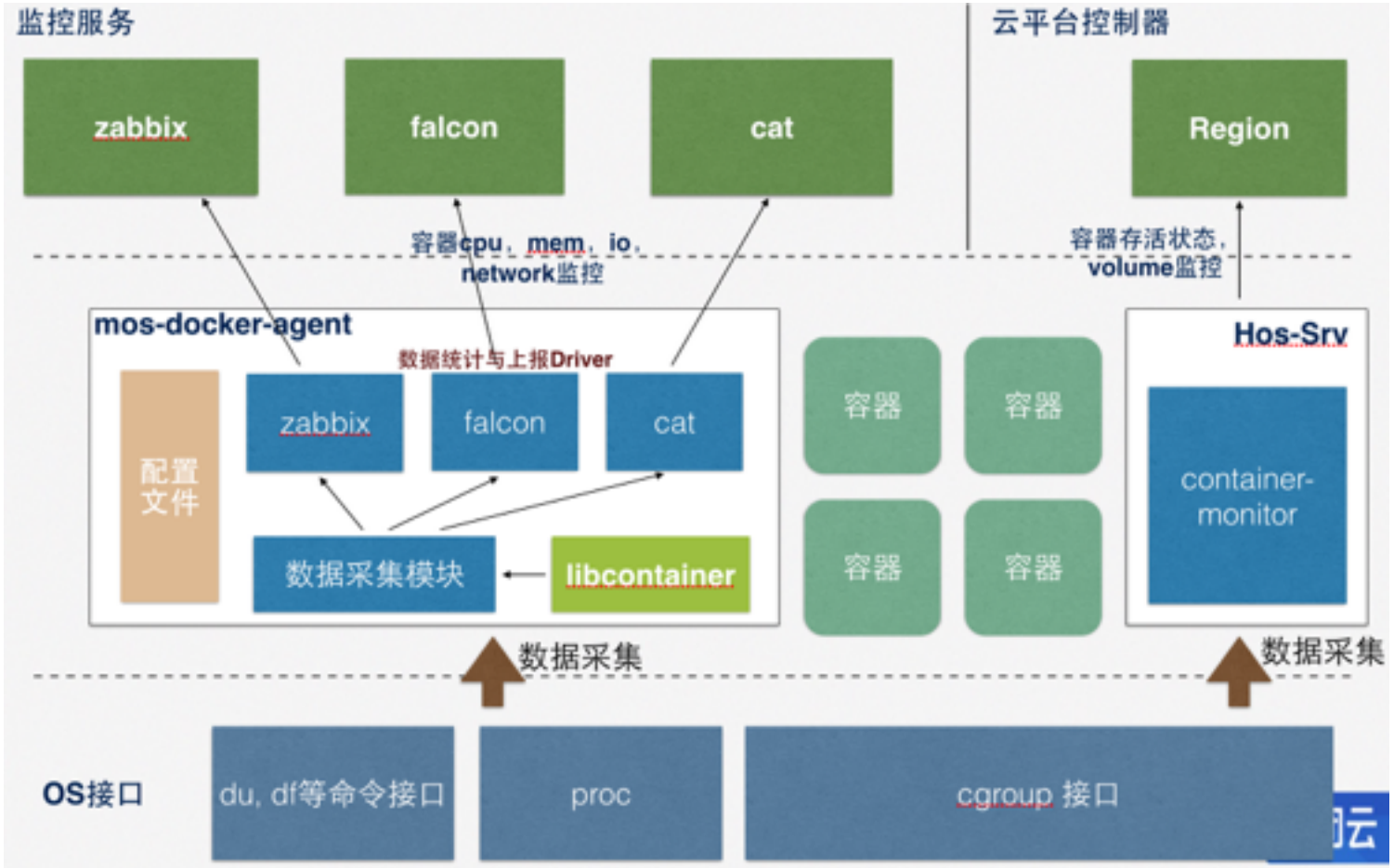
关键技术介绍

- 平台服务高可用性
- 容器伸缩高弹性
- 多层次的资源隔离技术
- 支持微服务类型业务

关键技术1：平台服务高可用性

- 高可用Docker平台控制器
  - nginx方案实现控制器多活，HA
  - 读写分离，性能保证
- 高可用容器服务
  - 支持容器服务以主备方式部署在不同宿主机上
  - 故障自动重建、自动迁移
  - volume备份，伪删除

强大的容器监控能力



- “双线”监控
- 完善的容器资源监控
- 兼容多种监控系统
- 简单，高效的监控配置

关键技术2：容器伸缩高弹性

- 水平伸缩
  - 数百个容器秒级扩容缩容
  - 支持容器依赖、互斥伸缩
  - 支持按宿主机打散、聚集伸缩
- 垂直伸缩
  - “不关机”扩容cpu、内存、volume

支持多种伸缩模式

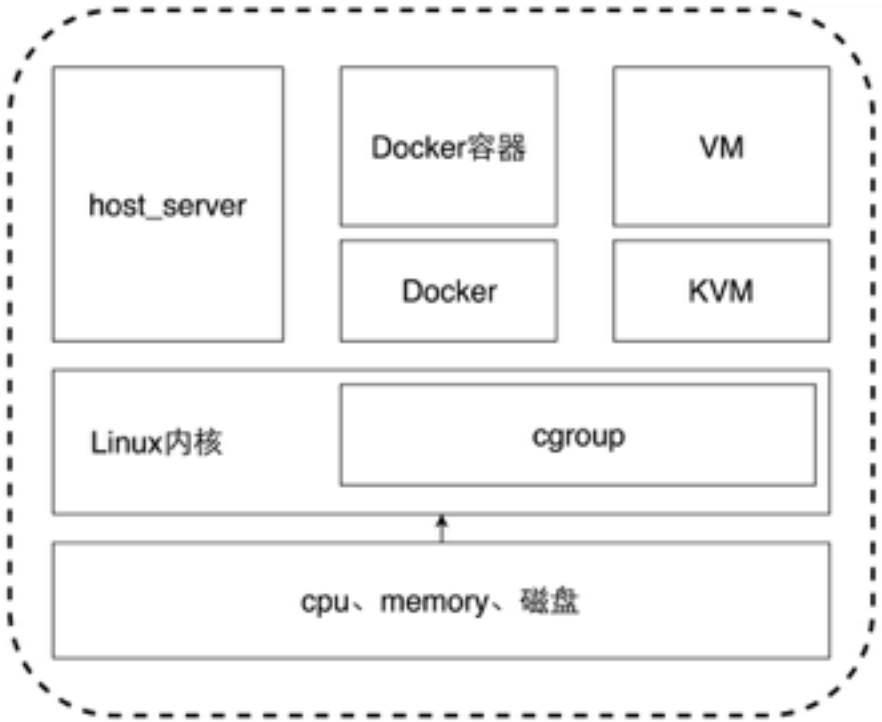
- 故障触发伸缩
  - 通过公司服务治理平台获取服务状态
  - 服务实例因故障退出后，自动创建新的服务实例
- 周期性伸缩
  - 可定制伸缩周期伸缩策略，使服务实例周期性伸缩
- 监控触发伸缩
  - 监控Docker平台CPU、存储和网络信息，通过自动伸缩保证资源合理分配

关键技术3：多层次的资源隔离

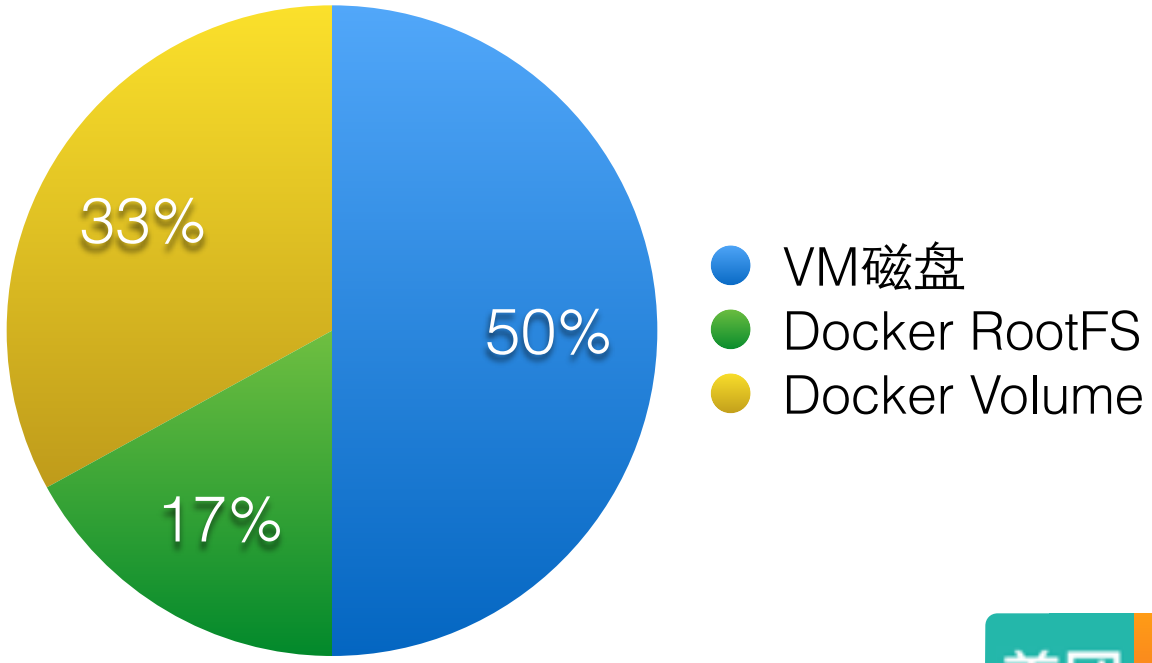
- 容器资源隔离
  - CPU、IO的软限制和硬限制：Cgroup
  - 内存：Cgroup
  - Docker volume磁盘容量：mount readonly（软限制），LVM（硬限制）
  - 网络带宽：Linux Traffic Control (TC)
- 宿主机隔离
  - 为业务提供宿主机资源池
- 多机房隔离
  - 提供业务多机房部署容灾能力

资源隔离之Docker / VM混部

统一的CPU隔离方案

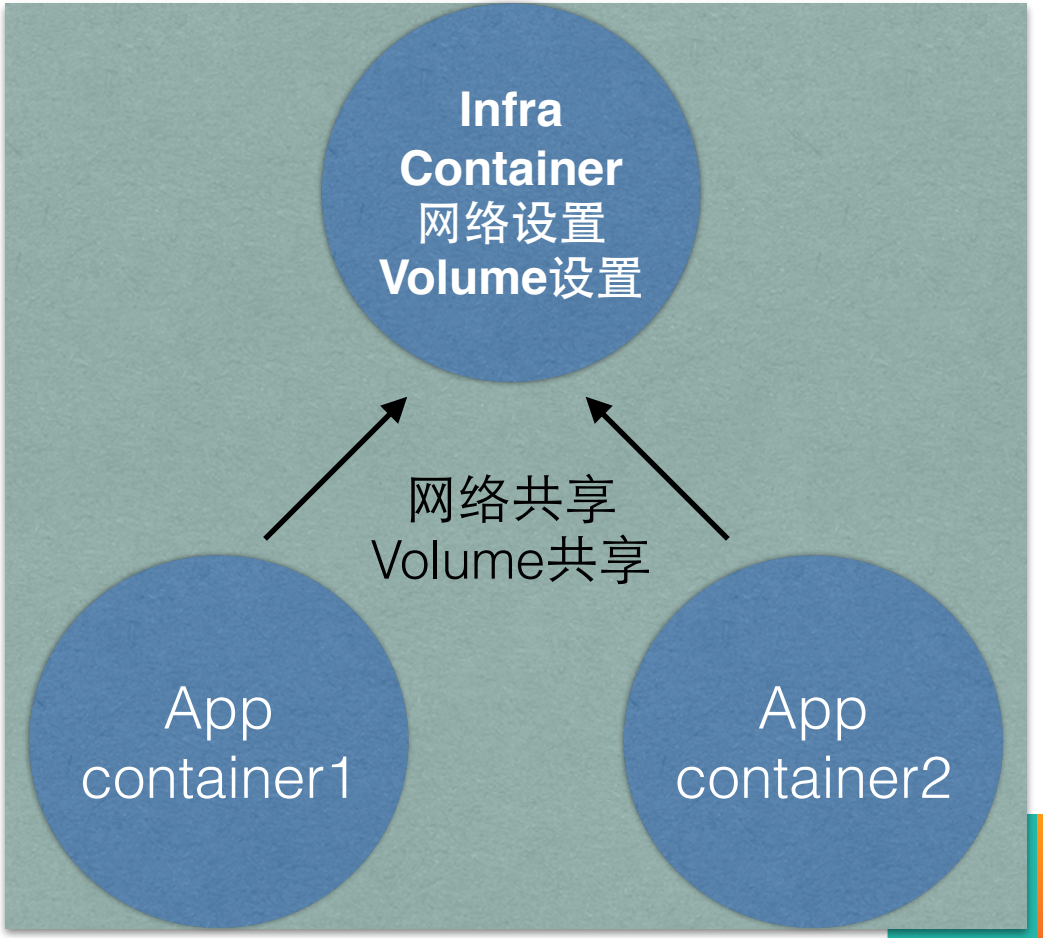


统一的磁盘管理



关键技术4：支持微服务类型业务

- set:
  - 所有容器都是以set方式运行
  - 一个业务实例
  - 一个或者多个容器组成
  - 调度的基本单位
  - 扩容缩容的基本单位
  - 多容器资源共享



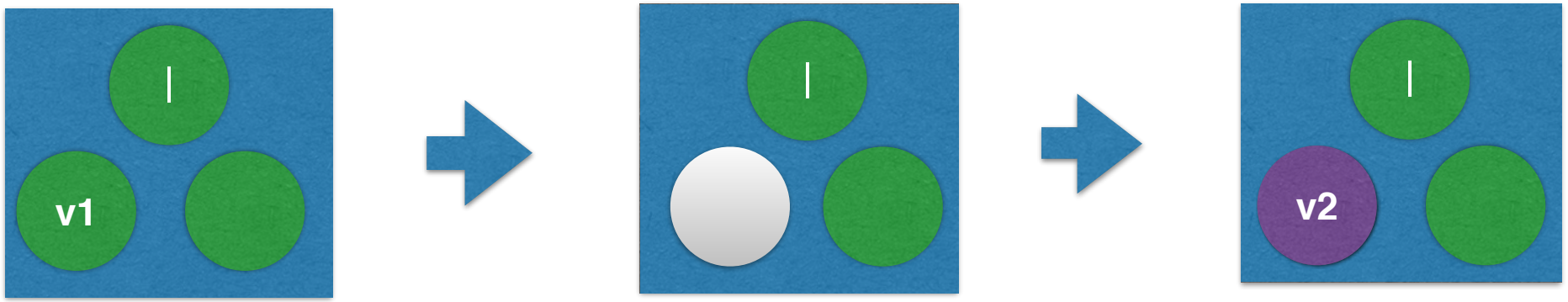
微服务特性介绍：运行时灰度更新

- 价值
  - 服务运行中更新组件
  - 秒级时延，服务几乎无中断
  - 服务容器集群批量更新

更新set metadata，
部署新版image

停止并删除对应容器

用新版image创建容器，
恢复网络和volume等设置



技术规划

	规划	说明
计算	Docker-in-KVM	满足多租户环境（公有云）对隔离性的要求
	容器Proc隔离	容器内看到的cpu、内存信息是容器的，而非宿主机的
存储	Volume使用LVM磁盘	支持Volume数据的容量限制
	Volume使用EBS块存储	支持Volume数据的高可靠性
网络	Docker美团云网络驱动	支持美团云网络架构的Docker网络驱动
	VPC	虚拟私有云

Q & A